MEMORIA
PROYECTO USISA
PREDICCIÓN DE VENTAS

Amalio Gómez López

1-Introducción

Esta memoria recopila los pasos que he llevado a cabo en el proyecto realizado para la empresa Unión Salazonera Isleña S.A., nombrada a partir de ahora como USISA.

Ubicada en Isla Cristina (Huelva), USISA ha liderado el sector alimentario, consolidándose como la conservera de pescado más grande en Andalucía.

Verticalmente integrada, USISA abarca todas las actividades, desde la producción hasta el servicio postventa.

Sergio Baeza-Herrazti, representante de USISA, busca un análisis temporal y una predicción a futuro de las ventas en la tienda online (www.usisa.com).

2- Datos

La empresa nos proporcionó todos los datos existentes de la tienda online. Aquí nos encontramos el primer obstáculo: en 2022 la empresa cambió de sistema de base de datos. Así que existen datos registrados en dos bases de datos diferentes:

Base de datos antigua:

Formato archivo: xlsx

Periodo: Julio 2017 - marzo 2022

Shape: 10769 filas× 7 columnas

Registra por fila cada producto del carrito de compra, por ejemplo si un cliente compra 5 productos distintos, el sistema registra 5 filas con el mismo id_order.

Columnas:
'id_order', 'Cliente', 'fecha', 'Referencia del pedido ', 'product_name', 'product_quantity', 'total_price_tax_incl'.

Base de datos nueva:

Formato archivo: csv

Periodo: Marzo 2022 - **noviembre 2023**

Shape: 2019 filas × 95 columnas

Registra todos los productos del carrito de compra, en una misma fila.

Columnas:
'order_id', 'order_number', 'order_date', 'paid_date', 'status', 'shipping_total', 'shipping_tax_total', 'fee_total', 'fee_tax_total', 'tax_total', 'cart_discount', 'order_discount', 'discount_total', 'order_total', 'order_subtotal', 'order_key', 'order_currency', 'payment_method', 'payment_method_title', 'transaction_id', 'customer_ip_address', 'customer_user_agent', 'shipping_method', 'customer_id', 'customer_user', 'customer_email', 'billing_first_name', 'billing_last_name', 'billing_company', 'billing_email', 'billing_phone', 'billing_address_1', 'billing_address_2', 'billing_postcode', 'billing_city', 'billing_state', 'billing_country', 'shipping_first_name', 'shipping_last_name', 'shipping_company', 'shipping_phone', 'shipping_address_1', 'shipping_address_2', 'shipping_postcode', 'shipping_city', 'shipping_state', 'shipping_country', 'customer_note', 'wt_import_key', 'shipping_items', 'fee_items', 'tax_items', 'coupon_items', 'refund_items', 'order_notes', 'download_permissions', 'meta:_wcpdf_invoice_number', 'meta:_wcpdf_invoice_date', 'meta:_wcpdf_invoice_number_data', 'meta:_wcpdf_invoice_date_formatted', 'meta:_wcpdf_invoice_settings', 'meta:_ppcp_paypal_fees', 'line_item_1', 'line_item_2', 'line_item_3', 'line_item_4', 'line_item_5', 'line_item_6', 'line_item_7', 'line_item_8', 'line_item_9', 'line_item_10', 'line_item_11', 'line_item_12', 'line_item_13', 'line_item_14', 'line_item_15', 'line_item_16', 'line_item_17', 'line_item_18', 'line_item_19', 'line_item_20', 'line_item_21', 'line_item_22', 'line_item_23', 'line_item_24', 'line_item_25', 'line_item_26', 'line_item_27', 'line_item_28', 'line_item_29', 'line_item_30', 'line_item_31', 'line_item_32', 'line_item_33'.

3-Objetivo

El objetivo del proyecto consiste en realizar una predicción de las ventas de la tienda online de USISA, abordando tanto el número de pedidos como la facturación total. Con los datos existententes se pueden obtener las siguientes variables:

> Total pedidos: Esta métrica refleja el número global de pedidos efectuados diariamente en la tienda online de USISA.
>
> Total Facturación: Esta variable se expresa en euros (€) y representa la suma total de ingresos generados a través de los pedidos realizados diariamente en la tienda online de USISA.
>
> Fecha: Indicador temporal que registra la fecha asociada a cada medición.

En cuanto a la duración del pronóstico futuro estará condicionada por el modelo seleccionado; en otras palabras, determinaremos el horizonte temporal que ofrezca los mejores resultados. Es importante tener en cuenta que a medida que aumentamos la cantidad de intervalos de tiempo pronosticados es probable que también incremente el margen de error. Por lo tanto, se buscará un equilibrio para lograr predicciones precisas sin comprometer la fiabilidad del modelo.

La hipótesis inicial sostiene que las ventas de USISA han experimentado un aumento significativo desde el lanzamiento de su tienda online en 2017, atribuyendo este crecimiento a la creciente popularidad del sitio web.

4-Limpieza y manipulación de datos

Para la predicción de clientes, hemos decidido prescindir de la información de productos en el dataframe antiguo (df_antiguo) debido a problemas de registro y falta de identificadores de producto. No obstante, planeamos incorporar detalles de productos en futuras predicciones utilizando exclusivamente el dataframe nuevo (df_nuevo), donde la información está completa y los productos cuentan con identificadores.

Actualmente no tenemos datos suficientes del dataframe nuevo, por lo que es fundamental contar con todos los datos históricos de la tienda online. De esta manera hemos fusionado el df_nuevo con el df_antiguo para tener una visión completa de la evolución a lo largo del tiempo.

Primero hemos limpiado el **df_nuevo**:

> Eliminación de columnas..
>
> Estandarizado de variables.
>
> Transformación de fecha a formato date time.

Selección de los pedidos con status completado.

Eliminación de pedidos de prueba.

Eliminación de NaNs.

Renombrar columnas.

Columnas df_nuevo después de limpieza:

'order_date', 'order_id', 'cliente_nombre', 'order_total'.

Limpieza **df_antiguo**:

Eliminación de columnas.

Estandarizado de variables.

Transformación de fecha a formato date time.

Agrupar pedidos por id_order.

Eliminación de pedidos de prueba.

Eliminación de NaNs.

Renombrar columnas.

Columnas df_antiguo después de limpieza:

'order_date', 'order_id', 'cliente_nombre', 'order_total'.

Después de limpiar el df_nuevo y el df_antiguo, hemos creado un nuevo data frame (**timeseries_diario**) donde concatenamos los dos data frames, agrupamos el número de pedidos y la facturación diariamente y eliminamos los saltos temporales.

Columnas timeseries_diario :

'total_pedidos', 'total_facturacion(€)'.

Indice df_cliente:

'fecha' (en formato date time).

5-Análisis exploratorio de los datos

Hemos realizado un análisis exploratorio de time series de frecuencia diaria de los datos de las columnas::

- *'total_pedidos'*
- *‘total_facturacion(€)’*

El EDA recoge:

- Visualización temporal: representación gráfica de la serie temporal del número de pedidos y de la facturación a lo largo del tiempo. Utiliza un gráfico de líneas para la facturación y de barras para el número de pedidos.
- Descomposición de la serie temporal: en componentes como tendencia, estacionalidad y residuos:
    - Tendencia: crecimiento positivo
    - Estacionalidad: True
    - Ruido: existe ruido al tener frecuencia diaria. Para el modelo se transformarán los datos a frecuencia mensual para reducir residuos.
- Autocorrelación: con gráficos de autocorrelación y autocorrelación parcial.

6- Predicción de ventas

6.1-Time Series

Para llevar a cabo la predicción futura, construimos un modelo de aprendizaje supervisado abordando los siguientes puntos clave:

Frecuencia de Datos:

- Después de probar con frecuencia diaria y semanal, optamos por una frecuencia mensual, ya que nuestro modelo ofrece mejores resultados con esta configuración.

Proporción de Datos en Train y Test:

- Seleccionamos los últimos 3 meses como conjunto de prueba y el resto para entrenamiento. Esta proporción demostró ser la más efectiva, proporcionando las métricas más sólidas para la generalización del modelo.

Horizonte de Predicción:

- Elegimos una frecuencia mensual y un horizonte de predicción de 3 meses, alineado con la muestra utilizada en el conjunto de prueba. Es esencial tener en cuenta que, a medida que aumentamos el horizonte de predicción, el error tiende a acumularse, siendo el último mes menos fiable.

Adicionalmente, excluimos los datos de noviembre debido a su incompletitud. Este enfoque nos permite adaptarnos a las particularidades de los datos y optimizar la capacidad predictiva del modelo.

6.2-AutoARIMA

Después de probar varios modelos de time series (RandomForestRegressor, XGBoostRegressor, ARIMA…). El modelo AutoARIMA que selecciona automáticamente los mejores parámetros para el modelo ARIMA, es el que mejores resultados me ha dado.

Resultados para AutoARIMA (facturación):

Mean Squared Error(MSE): 2764480.6480
Mean Absolute Error(MAE): 1191.4261
R^2 Score(r2): 0.0540

Resultados para AutoARIMA (pedidos):

Mean Squared Error(MSE): 665.8239
Mean Absolute Error(MAE): 23.6351
R^2 Score(r2): -9.81663400434568

6.3-Pycaret

Con el objetivo de lograr un modelo más eficaz, investigué y descubrí la biblioteca pycaret.time_series. Este enfoque sigue los siguientes pasos:

1. Evalúa todos los modelos de series temporales y los clasifica según su rendimiento.
2. Selecciona los mejores modelos.
3. Realiza un ajuste adicional de los parámetros para los modelos seleccionados.
4. Combina los mejores modelos para obtener un modelo mixto.
5. Realiza predicciones del modelo mixto en el conjunto de pruebas y proyecciones futuras.

Además, la biblioteca proporciona una función que visualiza los datos, muestra la separación entre el conjunto de entrenamiento y prueba, presenta las predicciones de los mejores modelos, muestra la predicción del modelo combinado y realiza proyecciones futuras. Este enfoque integral facilita la evaluación y el despliegue de modelos de series temporales de manera más eficiente.

Resultados para pycaret (facturación):

Mean Absolute Error(MAE): 1100.8623
R² Score(r2): 0.3662

Resultados para pycaret (pedidos):

Mean Absolute Error(MAE): 9.1750
R² Score(r2): -1.1216

6.4-Forecasting

Se utiliza el modelo mixto de pycaret para hacer una predicción de 3 meses a futuro.

2023-11: 20224.8027 €
2023-12: 22508.2809 €
2024-01: 10438.8487 €

2023-11: 170.9347 pedidos
2023-12: 177.9590 pedidos
2024-01: 94.7937pedidos

7-Conclusión

El objetivo principal se ha alcanzado de manera parcial. A pesar de la limitada cantidad de datos disponibles, hemos logrado desarrollar un sólido modelo de predicción para la facturación mensual a corto plazo. Sin embargo, en contraste, el modelo de predicción de pedidos mensuales no ha arrojado resultados favorables.

En resumen, hemos logrado desarrollar un modelo de forecasting capaz de realizar predicciones fiables para la facturación de uno o dos meses. No obstante, es importante señalar que a medida que intentamos predecir un mayor número de meses, se observa un aumento progresivo en el error.

En relación a la hipótesis, la confirmación de que la tendencia de las ventas de la tienda online de USISA es positiva desde su lanzamiento en 2017 fortalece nuestra evaluación. Además, nuestro modelo respalda y confirma esa tendencia de crecimiento para el futuro.

Adicionalmente, a medida que acumulemos datos suficientes en el futuro, podremos desarrollar un modelo exclusivamente con la base de datos nueva, permitiéndonos realizar predicciones a futuro por productos.